Questions on Topic Four: Exploring Bivariate Data

Multiple-Choice Questions

Directions: The questions or incomplete statements that follow are each followed by five suggested answers or completions. Choose the response that best answers the question or completes the statement.

1. As reported in *The New York Times* (September 21, 1994, page C10), a study at the University of Toronto determined that, for every 10 grams of saturated fat consumed per day, a woman's risk of developing ovarian cancer rises 20%. What is the meaning of the slope of the appropriate regression line?

   (A) Taking in 10 grams of fat results in a 20% increased risk of developing ovarian cancer.
   (B) Consuming 0 grams of fat per day results in a zero increase in the risk of developing ovarian cancer.
   (C) Consuming 50 grams of fat doubles the risk of developing ovarian cancer.
   (D) Increased intake of fat causes higher rates of developing ovarian cancer.
   (E) A woman's risk of developing ovarian cancer rises 2% for every gram of fat consumed per day.

2. A simple random sample of 35 world-ranked chess players provides the following statistics:

   Number of hours of study per day: $\bar{x} = 6.2$, $s_x = 1.3$
   Yearly winnings: $\bar{y} = \$208{,}000$, $s_y = \$42{,}000$
   Correlation $r = .15$

   Based on this data, what is the resulting linear regression equation?

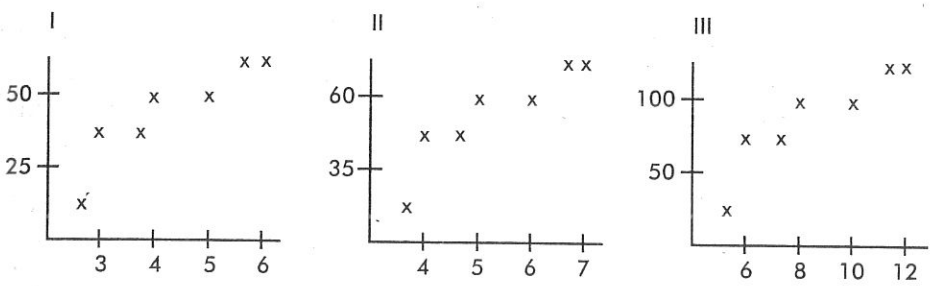
   (A) $\widehat{\text{Winnings}} = 178{,}000 + 4850\ \text{Hours}$
   (B) $\widehat{\text{Winnings}} = 169{,}000 + 6300\ \text{Hours}$
   (C) $\widehat{\text{Winnings}} = 14{,}550 + 31{,}200\ \text{Hours}$
   (D) $\widehat{\text{Winnings}} = 7750 + 32{,}300\ \text{Hours}$
   (E) $\widehat{\text{Winnings}} = -52{,}400 + 42{,}000\ \text{Hours}$

3. A scatterplot of a company's revenues versus time indicates a possible exponential relationship. A linear regression on Y = log(revenue in \$1000) against X = years since 2000 gives $\hat{y} = 0.67 + 0.82x$ with $r = .73$. Which of the following are valid conclusions?

   I. On the average, revenue goes up 0.82 thousand dollars (or \$820) per year.
   II. The predicted revenue in year 2005 is approximately 59 million dollars.
   III. 53% of the variation in revenue can be explained by variation in time.

   (A) I only
   (B) II only
   (C) III only
   (D) I and III
   (E) None of the above are valid conclusions.

4. Consider the following three scatterplots:



   Which has the greatest correlation coefficient?

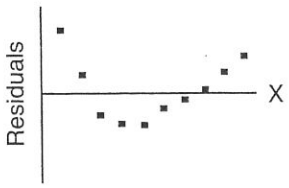
   (A) I
   (B) II
   (C) III
   (D) They all have the same correlation coefficient.
   (E) This question cannot be answered without additional information.

5. Suppose the correlation is negative. Given two points from the scatterplot, which of the following is possible?

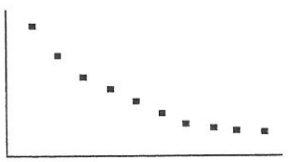
   I. The first point has a larger x-value and a smaller y-value than the second point.
   II. The first point has a larger x-value and a larger y-value than the second point.
   III. The first point has a smaller x-value and a larger y-value than the second point.

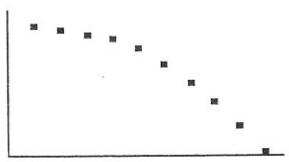
   (A) I only
   (B) II only
   (C) III only
   (D) I and III
   (E) I, II, and III

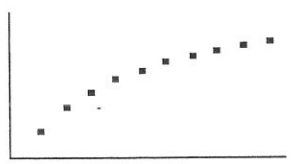
6. Consider the following residual plot:

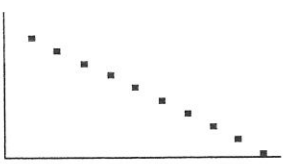


Which of the following scatterplots could have resulted in the above residual plot? (The y-axis scales are not the same in the scatterplots as in the residual plot.)

(A)


(B)


(C)


(D)


(E) None of these could result in the given residual plot.

7. Suppose the regression line for a set of data, $y = 3x + b$, passes through the point (2, 5). If $\bar{x}$ and $\bar{y}$ are the sample means of the x- and y-values, respectively, then $\bar{y}$ =

(A) $\bar{x}$.
(B) $\bar{x} - 2$.
(C) $\bar{x} + 5$.
(D) $3\bar{x}$.
(E) $3\bar{x} - 1$.

8. Su
to

I
II
III

(A)
(B)
(C)
(D
(E)

9. A s
for
bet

(A)
(B)
(C)

(D)

(E)

10. Wh

I
II

III

(A)
(B)
(C)
(D)
(E)

11. Wh

I.
II.
III.

(A)
(B)
(C)
(D)
(E)

8. Suppose a study finds that the correlation coefficient relating family income to SAT scores is $r = +1$. Which of the following are proper conclusions?

I. Poverty causes low SAT scores.
II. Wealth causes high SAT scores.
III. There is a very strong association between family income and SAT scores.

(A) I only
(B) II only
(C) III only
(D) I and II
(E) I, II, and III

9. A study of department chairperson ratings and student ratings of the performance of high school statistics teachers reports a correlation of $r = 1.15$ between the two ratings. From this information we can conclude that

(A) chairpersons and students tend to agree on who is a good teacher.
(B) chairpersons and students tend to disagree on who is a good teacher.
(C) there is little relationship between chairperson and student ratings of teachers.
(D) there is strong association between chairperson and student ratings of teachers, but it would be incorrect to infer causation.
(E) a mistake in arithmetic has been made.

10. Which of the following statements about the correlation r are true?

I. A correlation of .2 means that 20% of the points are highly correlated.
II. The square of the correlation measures the proportion of the y-variance that is predictable from a knowledge of x.
III. Perfect correlation, that is, when the points lie exactly on a straight line, results in $r = 0$.

(A) I only
(B) II only
(C) III only
(D) None of these statements is true.
(E) None of the above gives the complete set of true responses.

11. Which of the following statements about the correlation r are true?

I. It is not affected by changes in the measurement units of the variables.
II. It is not affected by which variable is called x and which is called y.
III. It is not affected by extreme values.

(A) I and II
(B) I and III
(C) II and III
(D) I, II, and III
(E) None of the above gives the complete set of true responses.

12. With regard to regression, which of the following statements about outliers are true?

    I. Outliers in the y-direction have large residuals.
    II. A point may not be an outlier even though its x-value is an outlier in the x-variable and its y-value is an outlier in the y-variable.
    III. Removal of an outlier sharply affects the regression line.

    (A) I and II
    (B) I and III
    (C) II and III
    (D) I, II, and III
    (E) None of the above gives the complete set of true responses.

13. Which of the following statements about influential scores are true?

    I. Influential scores have large residuals.
    II. Removal of an influential score sharply affects the regression line.
    III. An x-value that is an outlier in the x-variable is more indicative that a point is influential than a y-value that is an outlier in the y-variable.

    (A) I and II
    (B) I and III
    (C) II and III
    (D) I, II, and III
    (F) None of the above gives the complete set of true responses.

14. Which of the following statements about residuals are true?

    I. The mean of the residuals is always zero.
    II. The regression line for a residual plot is a horizontal line.
    III. A definite pattern in the residual plot is an indication that a nonlinear model will show a better fit to the data than the straight regression line.

    (A) I and II
    (B) I and III
    (C) II and III
    (D) I, II, and III
    (E) None of the above gives the complete set of true responses.

15. Data are obtained for a group of college freshmen examining their SAT scores (math plus writing plus critical reading) from their senior year of high school and their GPAs during their first year of college. The resulting regression equation is

$$\widehat{\text{GPA}} = 0.55 + 0.00161\ (\text{SAT total}) \quad \text{with} \quad r = .632$$

    What percentage of the variation in GPAs can be explained by looking at SAT scores?

    (A) 0.161%
    (B) 16.1%
    (C) 39.9%
    (D) 63.2%
    (E) This value cannot be computed from the information given.

Questions 16 and 17 are based on the following: The heart disease death rates per 100,000 people in the United States for certain years, as reported by the National Center for Health Statistics, were

Year:	1950	1960	1970	1975	1980
Death rate:	307.6	286.2	253.6	217.8	202.0

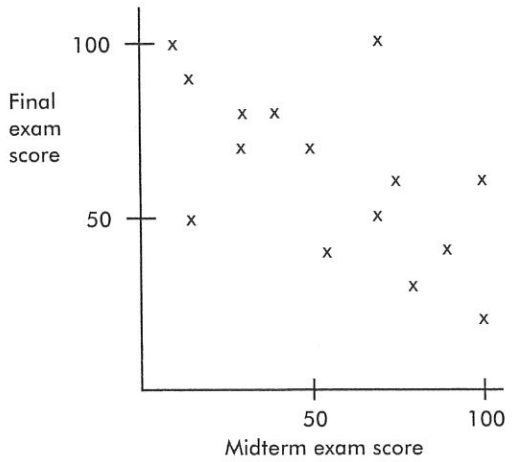
16. Which one of the following is a correct interpretation of the slope of the best-fitting straight line for the above data?

    (A) The heart disease rate per 100,000 people has been dropping about 3.627 per year.
    (B) The baseline heart disease rate is 7386.87.
    (C) The regression line explains 96.28% of the variation in heart disease death rates over the years.
    (D) The regression line explains 98.12% of the variation in heart disease death rates over the years.
    (E) Heart disease will be cured in the year 2036.

17. Based on the regression line, what is the predicted death rate for the year 1983?

    (A) 145.8 per 100,000 people
    (B) 192.5 per 100,000 people
    (C) 196.8 per 100,000 people
    (D) 198.5 per 100,000 people
    (E) None of the above

18. Consider the following scatterplot of midterm and final exam scores for a class of 15 students.



Which of the following are true statements?

I. The same number of students scored 100 on the midterm exam as scored 100 on the final exam.
II. Students who scored higher on the midterm exam tended to score higher on the final exam.
III. The scatterplot shows a moderate negative correlation between midterm and final exam scores.

(A) I and II
(B) I and III
(C) II and III
(D) I, II, and III
(E) None of the above gives the complete set of true responses.

19. If every woman married a man who was exactly 2 inches taller than she, what would the correlation between the heights of married men and women be?

(A) Somewhat negative
(B) 0
(C) Somewhat positive
(D) Nearly 1
(E) 1

20. Which of the following statements about the correlation r are true?

I. The correlation and the slope of the regression line have the same sign.
II. A correlation of –.35 and a correlation of +.35 show the same degree of clustering around the regression line.
III. A correlation of .75 indicates a relationship that is 3 times as linear as one for which the correlation is only .25.

(A) I and II
(B) I and III
(C) II and III
(D) I, II, and III
(E) None of the above gives the complete set of true responses.

21. Suppose the correlation between two variables is $r = .23$. What will the new correlation be if .14 is added to all values of the x-variable, every value of the y-variable is doubled, and the two variables are interchanged?

(A) .23
(B) .37
(C) .74
(D) –.23
(E) –.74

22. As reported in the *Journal of the American Medical Association* (June 13, 1990, page 3031), for a study of ten nonagenarians (subjects were age 90 ± 1), the following tabulation shows a measure of strength (heaviest weight subject could lift using knee extensors) versus a measure of functional mobility (time taken to walk 6 meters). Note that the functional mobility is greater with lower walk times.

Strength (kg):	7.5	6	11.5	10.5	9.5	18	4	12	9	3
Walk time (s):	18	46	8	25	25	7	22	12	10	48

What is the sign of the slope of the regression line and what does it signify?

(A) The sign is positive, signifying a direct cause-and-effect relationship between strength and functional mobility.
(B) The sign is positive, signifying that the greater the strength, the greater the functional mobility.
(C) The sign is negative, signifying that the relationship between strength and functional mobility is weak.
(D) The sign is negative, signifying that the greater the strength, the greater the functional mobility.
(E) The slope is close to zero, signifying that the relationship between strength and functional mobility is weak.

23. Suppose the correlation between two variables is –.57. If each of the y-scores is multiplied by –1, which of the following is true about the new scatterplot?

(A) It slopes up to the right, and the correlation is –.57.
(B) It slopes up to the right, and the correlation is +.57.
(C) It slopes down to the right, and the correlation is –.57.
(D) It slopes down to the right, and the correlation is +.57.
(E) None of the above is true.

24. A study of 100 elementary school children showed a strong positive correlation between weight and reading speed. Which of the following are proper conclusions?

I. Heavier elementary school children tend to have higher reading speeds.
II. Among elementary school children, faster readers tend to be heavier.
III. If you want to improve the reading speed of elementary school children, you should feed them more.

(A) I only
(B) I and II
(C) I, II, and III
(D) None of the three statements is a proper conclusion.
(E) None of the above gives the complete set of proper conclusions.

25. Consider the set of points {(2, 5), (3, 7), (4, 9), (5, 12), (10, n)}. What should n be so that the correlation between the x- and y-values is 1?

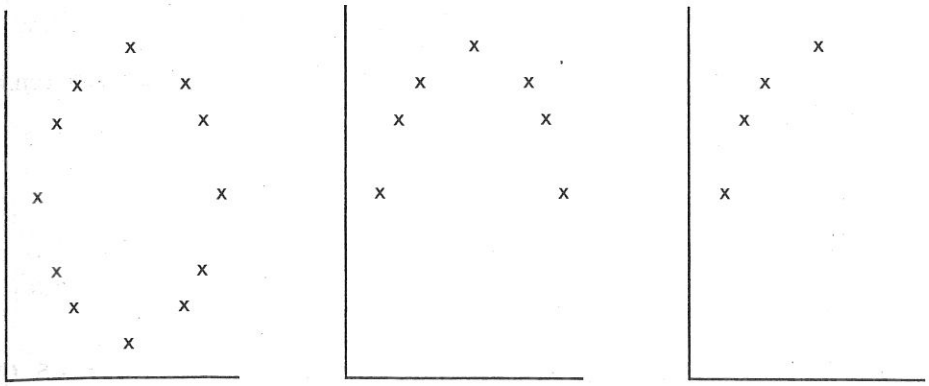
(A) 21
(B) 24
(C) 25
(D) A value different from any of the above.
(E) No value for n can make $r = 1$.

26. A study is conducted relating GPA to number of study hours per week, and the correlation is found to be .5. Which of the following are true statements?

I. On the average, a 30% increase in study time per week results in a 15% increase in GPA.
II. Fifty percent of a student's GPA can be explained by the number of study hours per week.
III. Higher GPAs tend to be associated with higher numbers of study hours.

(A) I and II
(B) I and III
(C) II and III
(D) I, II, and III
(E) None of the above gives the complete set of true responses.

27. Consider the following three scatterplots:



Which of the following is a true statement about the correlations for the three scatterplots?

(A) None are 0.
(B) One is 0, one is negative, and one is positive.
(C) One is 0, and both of the others are positive.
(D) Two are 0, and the other is 1.
(E) Two are 0, and the other is close to 1.

28. Consider the three points (2, 11), (3, 17), and (4, 29). Given any straight line, we can calculate the sum of the squares of the three vertical distances from these points to the line. What is the smallest possible value this sum can be?

(A) 6
(B) 9
(C) 29
(D) 57
(E) None of these values

29. Suppose that the scatterplot of log X and log Y shows a strong positive correlation close to 1. Which of the following is true?

I. The variables X and Y also have a correlation close to 1.
II. A scatterplot of the variables X and Y shows a strong nonlinear pattern.
III. The residual plot of the variables X and Y shows a random pattern.

(A) I only
(B) II only
(C) III only
(D) I and II
(E) I, II, and III

30. Which of the following statements about the correlation r are true?

I. When $r = 0$, there is no relationship between the variables.
II. When $r = .5$, 50% of the variables are closely related.
III. When $r = 1$, there is a perfect cause-and-effect relationship between the variables.

(A) I only
(B) II only
(C) III only
(D) I, II, and III
(E) All the statements are false.

31. Consider n pairs of numbers. Suppose $\bar{x} = 2$, $s_x = 3$, $\bar{y} = 4$, and $s_y = 5$. Of the following, which could be the least squares line?

(A) $\hat{y} = -2 + x$
(B) $\hat{y} = 2x$
(C) $\hat{y} = -2 + 3x$
(D) $\hat{y} = \frac{5}{3} - x$
(E) $\hat{y} = 6 - x$

Answer Key

1. **E**	8. **C**	14. **D**	20. **A**	26. **E**
2. **A**	9. **E**	15. **C**	21. **A**	27. **E**
3. **B**	10. **B**	16. **A**	22. **D**	28. **A**
4. **D**	11. **A**	17. **E**	23. **B**	29. **B**
5. **E**	12. **A**	18. **B**	24. **B**	30. **E**
6. **A**	13. **C**	19. **E**	25. **E**	31. **E**
7. **E**				

Answers Explained

1. **(E)** The slope is 20/10 = 2; that is, a woman's risk of developing ovarian cancer rises 2% for every gram of fat consumed per day. The other statements may be true, but they do not answer the question.

2. **(A)** Slope = $.15(\frac{42{,}000}{1.3}) \approx 4850$ and intercept = $208{,}000 - 4850(6.2) \approx 178{,}000$.

3. **(B)** log(revenue in \$1000), not revenue, goes up 0.82 per year.

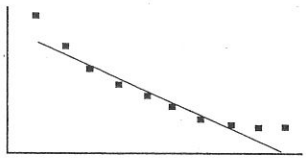
   log(revenue in \$1000) = 0.82(5) + 0.67 = 4.77 gives revenue = $10^{4.77}$ thousand dollars ≈ 59 million dollars.

   $r^2 = (.73)^2 = 53\%$ of the variation in log(revenue in \$1000), not revenue, can be explained by variation in time.

4. **(D)** The correlation coefficient is not changed by adding the same number to each value of one of the variables or by multiplying each value of one of the variables by the same positive number.

5. **(E)** A negative correlation shows a tendency for higher values of one variable to be associated with lower values of the other; however, given any two points, anything is possible.

6. **(A)** This is the only scatterplot in which the residuals go from positive to negative and back to positive.



7. **(E)** Since (2, 5) is on the line $y = 3x + b$, we have $5 = 6 + b$ and $b = -1$. Thus the regression line is $y = 3x - 1$. The point $(\bar{x}, \bar{y})$ is always on the regression line, and so we have $\bar{y} = 3\bar{x} - 1$.

8. **(C)** The correlation r measures association, not causation.

9. **(E)** The correlation r cannot take a value greater than 1.

10. **(B)** It can be shown that r^2, called the coefficient of determination, is the ratio of the variance of the predicted values $\hat{y}$ to the variance of the observed values y. Alternatively, we can say that there is a partition of the y-variance and that r^2 is the proportion of this variance that is predictable from a knowledge of x. In the case of perfect correlation, $r = \pm 1$.

11. **(A)** Correlation has the formula

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

We see that x and y are interchangeable, and so the correlation does not distinguish between which variable is called x and which is called y. The formula is also based on standardized scores (z-scores), and so changing units does not change the correlation. Finally, since means and standard deviations can be strongly influenced by outliers, the correlation is also strongly affected by extreme values.

12. **(A)** Removal of scores with large residuals but average x-values may not have a great effect on the regression line.

13. **(C)** An influential score may have a small residual but still have a greater effect on the regression line than scores with possibly larger residuals but average x-values.

14. **(D)** The sum and thus the mean of the residuals are always zero. In a good straight-line fit, the residuals show a random pattern.

15. **(C)** The coefficient of determination r^2 gives the proportion of the y-variance that is predictable from a knowledge of x. In this case $r^2 = (.632)^2 = .399$ or 39.9%.

16. **(A)** The regression equation is $\hat{y} = 7386.87 - 3.627x$, and so its slope is -3.627 and the y-value drops 3.627 for every unit increase in the x-value. Answer (C) is a true statement but doesn't pertain to the question.

17. **(E)** $-3.627(1983) + 7386.87 = 194.5$

18. **(B)** On each exam, two students had scores of 100. There is a general negative slope to the data showing a moderate negative correlation.

19. **(E)** On the scatterplot all the points lie perfectly on a line sloping up to the right, and so $r = 1$.

20. **(A)** The slope and the correlation are related by the formula

$$b_1 = r\frac{s_y}{s_x}$$

The standard deviations are always positive, and so b_1 and r have the same sign. Positive and negative correlations with the same absolute value indicate data having the same degree of clustering around their respective regression lines, one of which slopes up to the right and the other of which slopes down to the right. While $r = .75$ indicates a better fit with a linear model than $r = .25$ does, we cannot say that the linearity is threefold.

21. **(A)** The correlation is not changed by adding the same number to every value of one of the variables, by multiplying every value of one of the variables by the same positive number, or by interchanging the x- and y-variables.

22. **(D)** The slope is negative (–2.4348); that is, the regression line slopes down to the right, indicating that nonagenarians with greater strength have lower walk times and thus greater functional mobility.

23. **(B)** The slope and the correlation coefficient have the same sign. Multiplying every y-value by –1 changes this sign.

24. **(B)** Correlation shows association, not causation. In this example, older school children both weigh more and have faster reading speeds.

25. **(E)** A scatterplot readily shows that while the first three points lie on a straight line, the fourth point does not lie on this line. Thus no matter what the fifth point is, all the points cannot lie on a straight line, and so r cannot be 1.

26. **(E)** Only III is true. A positive correlation indicates that higher values of x tend to be associated with higher values of y.

27. **(E)** All three scatterplots show very strong nonlinear patterns; however, the correlation r measures the strength of only a linear association. Thus $r = 0$ in the first two scatterplots and is close to 1 in the third.

28. **(A)** Using your calculator, find the regression line to be $\hat{y} = 9x - 8$. The regression line, also called the least squares regression line, minimizes the sum of the squares of the vertical distances between the points and the line. In this case (2, 10), (3, 19), and (4, 28) are on the line, and so the minimum sum is $(10 - 11)^2 + (19 - 17)^2 + (28 - 29)^2 = 6$.

29. **(B)** When transforming the variables leads to a linear relationship, the original variables have a nonlinear relationship, their correlation (which measures linearity) is not close to 1, and the residuals do not show a random pattern.

30. **(E)** These are all misconceptions about correlation. Correlation measures only linearity, and so when $r = 0$, there still may be a nonlinear relationship. Correlation shows association, not causation.

31. **(E)** The least squares line passes through $(\overline{x}, \overline{y}) = (2, 4)$, and the slope b satisfies $b = r\frac{s_y}{s_x} = \frac{5r}{3}$. Since $-1 \le r \le 1$, we have $-\frac{5}{3} \le b \le \frac{5}{3}$.